METHOD FOR GENERATING TRAINING DATA FOR MEDICAL TEXT ABBREVIATION AND ACRONYM NORMALIZATION

Abstract

A method for electronically generating high-quality feature vectors that can be used in connection with electronic data processing systems implementing Maximum Entropy or other statistical models to accurately normalize abbreviations in text such as medical records. An abbreviation database and a training text database are provided. The abbreviation database includes abbreviation data representative of abbreviations and associated expansions to be normalized. The training text database includes a corpus of text having expansions of the abbreviations to be normalized. The corpus of text is processed as a function of the abbreviation data to identify the expansions in the corpus of text. Context information describing the context of the text in which the expansions were identified is generated. A set of feature vectors is also stored. Each feature vector including the context information generated for the associated expansion identified in the corpus of text.

M2:20554317.01